Genome-wide detection and family clustering of ion channels

Rachel Harte, Christos A. Ouzounis*

Computational Genomics Group, Research Programme, The European Bioinformatics Institute, EMBL Cambridge Outstation, Cambridge CB10 1SD, UK

Received 16 July 2001; revised 30 November 2001; accepted 30 November 2001

First published online 11 December 2001

Edited by Gunnar von Heijne

Abstract Ion channels represent an important class of molecules that can be classified into 13 distinct groups. We present a strategy using a 'learning set' of well-annotated ion channel sequences to detect homologues in 32 entire genome sequences from Archaea, Bacteria and Eukarya. A total of 299 putative ion channel protein sequences were detected, with significant variations across species. The clustering of these sequences reveals complex relationships between the different ion channel families.

Key words: Ion channels; Genome analysis; Sequence clustering; Protein families

1. Introduction

Ion channels constitute a large class of extensively studied proteins, but as yet there has not been an analysis of ion channels present in completed genome sequences, nor have the sequence relationships of these families been investigated simultaneously. Also, many members of this class encoded in entire genomes have not yet been recognised as ion channel proteins. The availability of a number of genomes from all three domains of life – Archaea, Bacteria and Eukarya – now provides the opportunity for a genome-wide study of this important class of proteins.

Ion channels are responsible for facilitating selective uptake or release of ions from the cell. They are therefore involved in a variety of functions from signal transmission in triploblastic nervous systems to modulation of swimming behaviour in *Paramecium* sp. [1]. They are integral membrane proteins which allow for fast transmembrane passive diffusion flow of selected inorganic ions via a hydrophilic pore at around 10 000 000 ions/channel/s [1].

Known ion channel families can be classified into 13 groups according to their structural, functional and physiological properties:

1. Neurotransmitter (extracellular ligand-gated) – includes nicotinic acetylcholine and serotonin (5-hydroxytryptamine) receptors (excitatory), and glycine and γ-butyric acid receptors (inhibitory) [2]
2. Ionotropic glutamate-activated cationic channels (iGluR) (extracellular ligand-gated) – includes kainate, α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid and *N*-methyl-D-aspartate receptors [3,4]
3. Cyclic nucleotide-gated (CNG) or intracellular ligand-gated [5,6]
4. Voltage-gated potassium (including calcium-sensitive maxi K_{Ca} channels) [7]
5. Voltage-gated sodium [8]
6. Voltage-gated calcium [8]
7. Voltage-gated chloride (CLC) [9,10]
8. Inward rectifier potassium (IRK) [11,12]
9. Two-pore potassium [13–15]
10. Potassium channels in prokaryotes [16]
11. ATP-gated purinergic (P2X) [17,18]
12. Degenerin/amiloride-sensitive sodium epithelial (ASC) [17,19]
13. Large-conductance mechanosensitive (MscL) [17,20].

Channels may be regulated by phosphorylation such as the sodium ASC [21], voltage-gated sodium channels [22] and the P2X channels [23]. Charged groups at either end of the pore can aid selectivity [3] and channels can exist either in an open or closed state while transition to the open state may be triggered by the binding of an agonist or a voltage change across the membrane. There is a pore segment of the protein which is membrane-associated and forms the pore (P) region which acts as a selectivity filter in ion channels (Fig. 1). Functional diversity of ion channels is achieved by alternative splicing in many channel types including potassium channels [7,11,12,24], sodium channels [25], CNG channels [26] and iGluR channels [27]. Variation in functional and physiological properties may also be achieved by the formation of homomultimeric or heteromultimeric assemblies of subunits. Recently, the structure for the P region of the potassium channel (KcsA) from *Streptomyces lividans* was solved [28] and this represents a significant and exciting advance in ion channel research. It has led

*Corresponding author. Fax: (44)-1223-494471.
E-mail address: ouzounis@ebi.ac.uk (C.A. Ouzounis).

PII: S0014-5793(01)03254-9

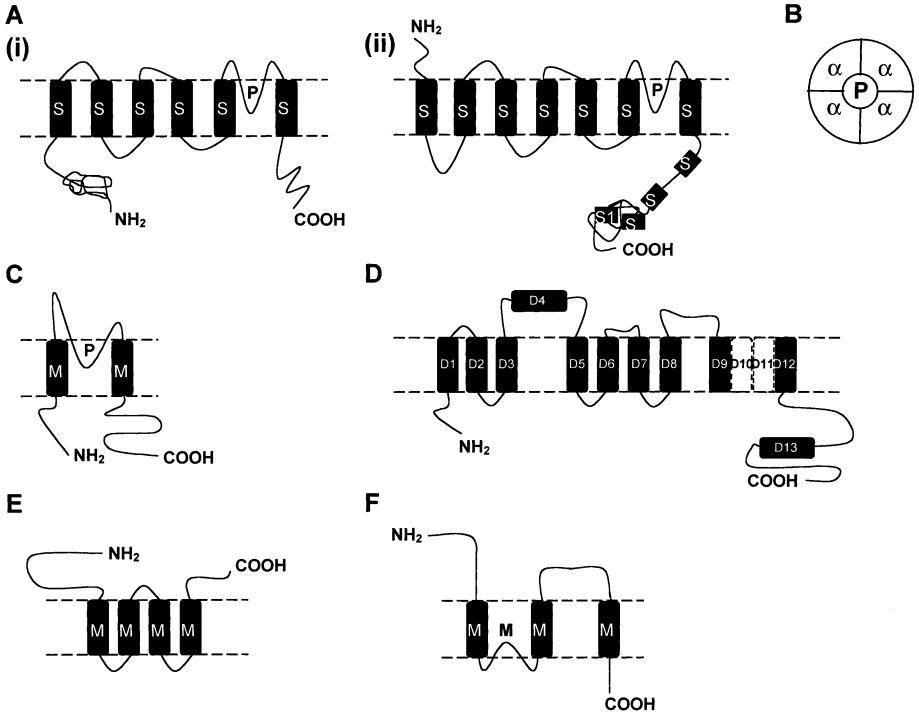

Fig. 1. Schematic representation of the structure for members of various ion channel families. Black barrels represent transmembrane domains. A: Voltage-gated potassium channels. Positively charged residues in S4 sense voltage. (i) α subunit of voltage-gated potassium channels. Voltage-gated sodium and calcium channels have four domains in one polypeptide chain where each domain is similar in sequence and structure to the α subunit of voltage-gated potassium channels. CNG channels show structural homology to these channels. (ii) α subunit of maxi K_{Ca} voltage-gated potassium channels. B: Channel of four subunits showing how the P region of α subunit of a voltage-gated potassium channel or domain of the voltage-gated sodium and calcium channels contribute to the pore part of the channel. Other channels have a similar structure composed of differing numbers of subunits. C: IRK – M1 and M2 are homologous in sequence and structure to S5 and S6 of voltage-gated potassium channels [29]. IRK channels – Kir1.0 and Kir6.0 subfamilies have ATP-binding regions (inhibitory, involve both N- and C-termini binding sites) and Kir channels can be modulated by PIP_2 (C-terminus binding sites) as well as by $G_{\beta\gamma}$ (N- and C-termini binding sites) in the case of Kir3.0. Kir6.0 family members have associated membrane-bound sulphonylureas which are involved in their regulation [12]. Prokaryotic potassium channels, ASC and MscL channels all have a similar structure. D: CLC – The white transmembrane regions are ones that may not exist as it is difficult to determine the exact number of domains in this region. E: Neurotransmitter-gated channel – M2 contributes mostly to the P region. P2X channels also have four transmembrane regions with a small P region next to M2, and in addition they have a large cysteine-rich extracellular loop containing a phosphate-binding domain for ATP. F: iGluR – The P region is M2. Ligands bind to extracellular loops.

to a greater understanding of ion selectivity by the P region of these channels [28].

Few of these ion channels have been identified in Bacteria or Archaea, so an analysis of these genomes may reveal some new putative ion channel proteins both in these groups and in the Eukarya. It is important to understand these proteins, as they are involved in some of the most fundamental aspects of life such as transmission of nerve signals and muscle function. Mutation of these proteins can also be the cause of many diseases [29][1] so knowledge of ion channels and their structural and functional relationships may lead to a better understanding of the pathogenesis of diseases in which they are involved.

2. Materials and methods

Here, we describe the distribution of ion channels found in a number of completed genomes. Ion channel families were identified using the clustering programme, GeneRAGE [30], which automatically generates clusters of related sequences and identifies the multi-domain structure of proteins. Genomes analysed included six from Archaea, 23 from Bacteria and three from Eukarya[2]. Profile Hidden Markov Models (HMMs) [31] were built for each family and used in an iterative search against the genome sequences to further detect novel members of those families as a sensitive method of identifying remote homologues.

A 'learning set' of 48 query sequences were assembled from the literature and from the InterPro database[3] for each known ion channel family. Using BLASTp [32], a search was performed against Swiss-Prot for each of these sequences to identify further family members which were well-annotated. A total of 292 proteins in this 'seed set' were then used to detect 299 putative ion channel sequences in complete genomes using BLASTp with an *E*-value threshold of 10^{-8}. The CAST filtering algorithm [33] was employed to mask compositionally biased regions in order to increase the sensitivity of this BLASTp search. GeneRAGE was used to cluster these proteins together with 48 of the sequences identified in the 'learning set' of Swiss-Prot sequences, which were then used to annotate the clusters. An *E*-value threshold of 10^{-4} was used for the all-against-all BLASTp analysis subsequently used for the full clustering of the putative ion channel sequences by GeneRAGE [30]. Homology relationships were visualised using the graph layout optimisation algorithm biolayout [34].

[1] For a review of diseases caused by ion channel mutations, see http://www.neuro.wustl.edu/neuromuscular/mother/chan.html.

[2] Details of the genomes studied and their protein sequences may be viewed at http://www.ebi.ac.uk/research/cgg/discovery/ionc.html

[3] http://www.ebi.ac.uk/interpro/

HMMs were built from multiple alignments created with CLUSTALX using hmmbuild and hmmcalibrate from the HMMER (version 2.1.1) package of programmes[4]. hmmsearch used the profile HMMs to search against the genome protein sequences. No additional putative ion channels were found for the P2X, CNG (plus ether-à-go-go or eag) and voltage-gated cationic ion channel families using HMMs to search the target genome sequences. The HMM-based sequence similarity search provided 1–15 more putative ion channels for each of the other families resulting in a total of 39 additional channels found.

3. Results

3.1. Family clustering of ion channels

GeneRAGE identified six main clusters, of which cluster II (Table 1) includes a number of the ion channel families (Fig. 2), while the remaining five clusters correspond to five classes (Table 1). The large, complex cluster was further resolved using biolayout [34]. These protein sequences are highly similar and contain many multi-domain proteins. It was not possible to resolve the cation specificity of the voltage-gated cation channels (potassium, sodium, calcium), because of the extensive degree of sequence similarity within this group (Fig. 2). Voltage-gated potassium channels consist of four α subunits, each of which resembles the four domains of sodium and calcium channels in structure (Fig. 1), and these are thought to have arisen through gene duplication of potassium channels [8,16,22,25,29,35].

It is notable that the eag family members cluster with the CNG channels but show also similarity to the potassium channels, indicating that these are multi-domain proteins. A search against Pfam [36] revealed that these two families both contain the CNG channel domain and a cyclic nucleotide-binding domain and it may be that one group is the evolutionary precursor of the other. Eag family members are usually classed as voltage-gated potassium channels based on their biochemical and physiological properties, and they are thought to be most closely related to CNG and IRK channels [37,38].

Further analysis of the output of biolayout (Fig. 2) revealed that there is a multi-domain protein with two domains related to potassium channels and iGluRs, confirming previous observations [39]. This protein, GluR0, is an unusual glutamate-gated channel being highly selective for potassium ions and which is the only one so far (including the present study) to be found in prokaryotes. This discovery adds evidence to the hypothesis that iGluRs may have existed before the major eukaryotic divisions occurred [4]. The P region of iGluRs is also homologous to that found in voltage-gated potassium, sodium and calcium channels as well as in IRK and CNG channels [4,40]. iGluRs are thought to contain a periplasmic-binding protein-like domain and are similar to bacterial glutamine-binding proteins [4,40]. Bacterial ion channels containing this type of domain were found (with lower scores than the iGluRs) when using the iGluR HMM profile to search complete genome sequences, so they probably represent distant relatives of this family which has evolved to have a more specialised function.

The IRK family formed a distinct cluster except that IRK2 from rat was found to be a multi-domain containing an IRK family domain and another similar to KcsA (bacterial potassium channel, Swiss-Prot identifier: KCSA_STRLI). From an analysis of the P region, IRK channels form a family distinct from voltage-gated and other potassium channels with their closest relatives being the *slo* channel gene and the KchC gene from *Archaeoglobus fulgidus* and the KchA gene of *Escherichia coli* [16]. The large cluster produced by GeneRAGE included these potassium channels and therefore other channels related to these with the IRK family proteins.

Table 1
Clusters of ion channel proteins produced by GeneRAGE

Cluster	Ion channel protein families represented
I	CLC
II	voltage-gated potassium (including maxi K_{Ca})
	voltage-gated sodium
	voltage-gated calcium
	IRK
	two-pore potassium
	prokaryotic potassium
	iGluR
III	neurotransmitter-gated
IV	ASC
V	P2X
VI	MscL

A search of Pfam showed that some of the potassium channels identified in the Bacteria and Archaea contain domains similar to the TrkA potassium channel as well as an NAD-binding domain and in the *Synechocystis* sp. potassium channel there were found to be two NAD-binding domains. It is thought that potassium channels with an NAD-binding domain may have developed only in the prokaryotes and these may be close to an ancestral form of potassium channels [16]. It has been suggested that voltage-gated potassium channels were the evolutionary precursors of the other voltage-gated cation channels [16,25]. Support for this hypothesis also comes from the fact that potassium channels seem to appear in the Bacteria before voltage-gated channels evolved.

Thus, the major cluster (listed as cluster II in Table 1) contains representatives from eight groups (see Section 1), strongly suggesting extensive evolutionary relationships and a common ancestral state. The remaining five groups (neurotransmitter, CLC, P2X, ASC, MscL) appear to be confined to a specific protein family, each without detectable sequence similarity to any of the other groups (Fig. 2).

Ivermectin-sensitive glutamate-gated chloride channels found in *Caenorhabditis elegans* have previously been thought to have no vertebrate homologues [41]. A BLASTp similarity search of ion channel proteins from the genomes against nrdb90 [42] revealed proteins which were highly similar to these channels. These putative ivermectin-sensitive glutamate-gated chloride channels formed a cluster with the neurotransmitter-gated channels (including some vertebrate channels from the seed set). This suggests that they are remote homologues of the neurotransmitter-gated ion channels and not related to the glutamate-gated ion channel family (Fig. 2).

3.2. Phylogenetic distribution of ion channels

There is a progressive increase in ion channel families up the evolutionary scale with respect to the genomes of the species analysed. In the Archaea, only CLC or potassium channels were found and, interestingly, every species has at least one ion channel. In the Bacteria, apart from the above-

[4] http://hmmer.wustl.edu

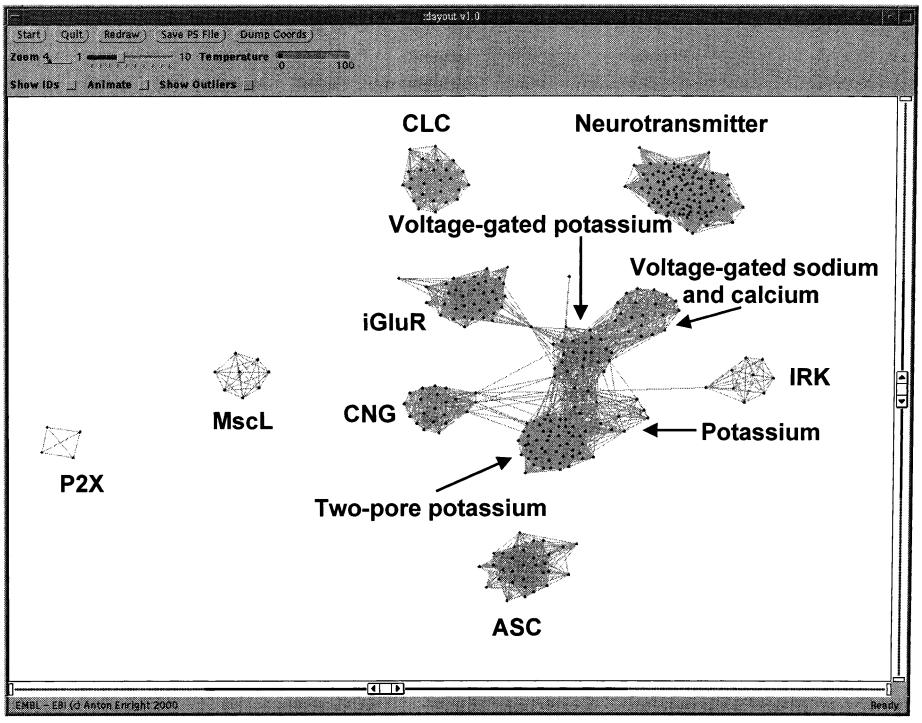

Fig. 2. Sequence similarities within and between major ion channel families. The graph has been produced using biolayout [34]. Vertices (points) represent proteins and the edges (connecting lines) represent homology relationships detected by GeneRAGE.

mentioned two families, there is a glutamate-gated channel (GluR0) which has been previously identified in the cyanobacterium *Synechocystis* sp., and members of the MscL family. Bacteria and Archaea do not appear to have evolved a wide range of proteins belonging to the ion channel class, but they do have many other ion transport proteins that do not belong to this class, but perform a similar function [43].

Only six species (all in the Bacteria) have the MscL channel and in *Haemophilus influenzae* this is the only ion channel present. Eleven out of the 23 species of Bacteria have no channels detectable by sequence similarity to the seed set: *Borrelia burgdorferi*, *Campylobacter jejuni*, *Chlamydia pneumoniae* AR39, *C. pneumoniae* CWL029, *Chlamydia trachomatis* serovar D, *C. trachomatis MoPn* Nigg, *Mycoplasma genitalium*, *Mycoplasma pneumoniae*, *Rickettsia prowazekii*, *Treponema pallidum*, and *Ureaplasma urealyticum*. It is intriguing that all these species are pathogenic. It has been observed that pathogenic bacteria often express a reduced set of proteins compared to free-living bacteria which is probably due to gene loss as they are more dependent on host mechanisms. *H. influenzae* and particularly *Mycoplasma* sp. are missing many inorganic ion transport proteins, the latter species hav-

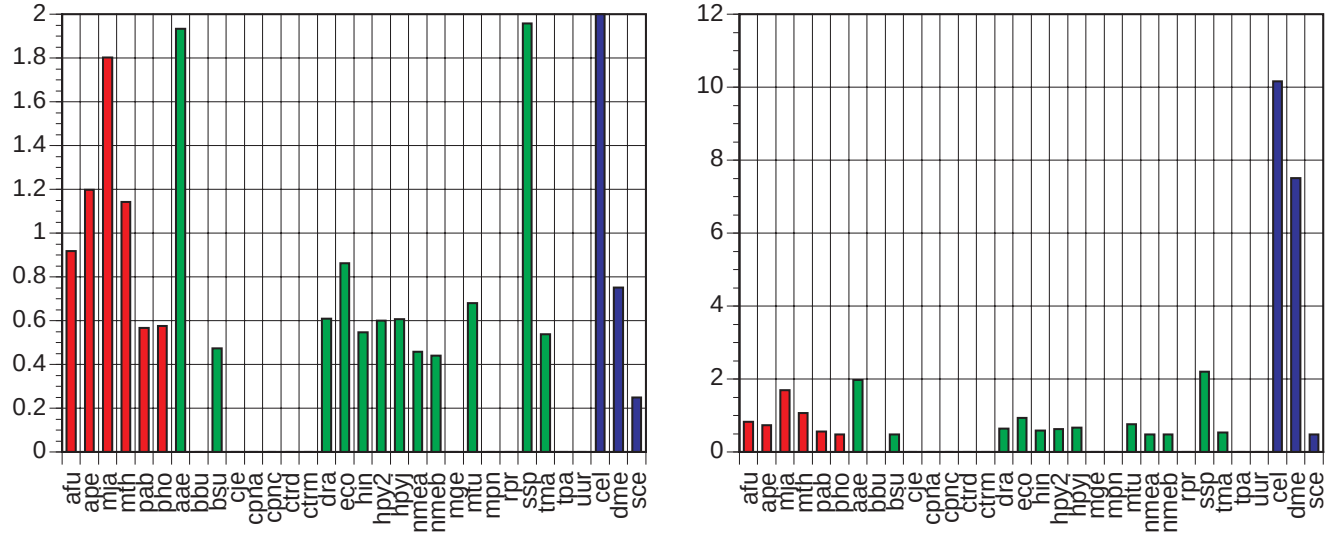

Fig. 3. Number of ion channels normalised for (left) the number of open reading frames for the genome of each species and (right) genome size. Key: red – Archaea; green – Bacteria; blue – Eukarya. Species names abbreviations available on the web site. Axes: *x*-axis represents individual species, *y*-axis represents counts for ion channels per (left) per 1000 ORFs and (right) per megabase.

ing particularly reduced genomes. An exception to this is the pathogenic bacterium, *Mycobacterium tuberculosis*, whose genome appears to encode three ion channel proteins. A more thorough survey may be conducted when genomes of more non-pathogenic, free-living bacterial genomes become available.

With the evolution of the Eukarya, there seems to be a large increase in the diversity of ion channels in the eukaryotes and also an apparent loss of the MscL channels. *Saccharomyces cerevisiae* has only three ion channels which were identified by a BLASTp search: the CLC channel GEF1, the voltage-gated calcium channel CCH1 and also the two-pore potassium channel TOK1. It thus appears that voltage-gated calcium channels originated before the appearance of this species. The two-pore potassium channel TOK1 (8TM regions) has no counterparts in any other genome in this study while *Drosophila melanogaster* and *C. elegans*, in particular, possess a large number of the related TWIK type channels (4TM regions).

Of the Eukarya, *C. elegans* (194 ion channels) and *D. melanogaster* (103 ion channels) have the most channels of the genomes studied and these include representatives of all the ion channel families except for MscL. *D. melanogaster* has one possible maxi K_{Ca} channel which appears to be *Dslo*. *S. cerevisiae* has the lowest number of ion channels when normalised for either genome size or number of open reading frames (Fig. 3) so complexity is not the only factor involved in the possession of a large number and diversity of ion channels. It is interesting that there is a huge number of ion channels appearing not only with the increase in complexity in the higher eukaryotes (*C. elegans* and *D. melanogaster*), but maybe also with the evolution of nervous systems whose functions involve many ion channels. Compared to *Drosophila*, *C. elegans* has much larger numbers of neurotransmitter-gated channels and also more TWIK-related two-pore potassium channels. It is thought that the unusually large diversity of neurotransmitter-gated channels in this species is due to the fact that channels may be localised to particular synapses or regions of the cell [41]. It is also possible that incorrect gene prediction in these species may have led to an underestimation in the numbers of these channels. Consistent with previous observations, no voltage-gated sodium channels have been found in *C. elegans* [41] but there is one detected in *D. melanogaster* (paralytic protein) which is the only one identified in this study. Both species have ASCs.

Interestingly, no ATP-gated channels were found in any genome studied and all those found in Swiss-Prot are from mammals. It may be that this family of proteins only evolved in mammals but this remains to be confirmed as more genomes are sequenced. The Kir6 subfamily of IRK channels also contain an ATP-binding domain [12] and no channels showing similarity to these proteins were identified in the genomes studied. It may be that this type of channel also appeared late in evolution. As more genome sequences are completed it will be possible to see whether this is the case and if these ATP-binding regions are related.

In conclusion, we were able to rapidly cluster large and complex multi-domain proteins as found in the ion channel protein class and delineated their family relationships. From the clusters formed, together with similarity searches performed, it appears that there were no false positives in this study, but it is possible that some very remote homologues have not been detected. Consideration of the habitat and physiology of the organisms can provide explanations for the presence or absence of certain families of ion channels. Further searches for ion channels as more genomes are sequenced may reveal more intricate evolutionary patterns of this interesting class of protein molecules.

Acknowledgements: The authors thank Declan Doyle (Department of Biochemistry, University of Oxford), Anton Enright (Computational Genomics Group, EBI) and Iain Paulsen (The Institute for Genome Research) for comments. Part of this work was submitted to the University of York for the MRes degree in Bioinformatics (R.H.). C.A.O. acknowledges support by EMBL, the European Commission (DGXII – Science, Research and Development), the UK Medical Research Council and IBM Research.

References

[1] Miller, C. (1991) Science 252, 1092–1096.
[2] Ortells, O. and Lunt, G.G. (1995) Trends Neurosci. 18, 121–127.
[3] Unwin, T. (1993) Cell 72, 31–41.
[4] Chiu, J., DeSalle, R., Lam, H.-M., Meisel, L. and Coruzzi, G. (1999) Mol. Biol. Evol. 16, 826–838.
[5] Zufall, F., Shepherd, G.M. and Barnstable, C.J. (1997) Curr. Opin. Neurobiol. 7, 404–412.
[6] Scott, S.-P., Cummings, J., Joe, J.C. and Tanaka, J.C. (2000) Biophys. J. 78, 2321–2333.
[7] Toro, L., Wallner, M., Meera, P. and Tanaka, Y. (1998) News Physiol. Sci. 13, 112–117.
[8] Caterall, W.A. (1995) Annu. Rev. Biochem. 64, 493–531.
[9] Jentsch, T.J. (1996) Curr. Opin. Neurobiol. 6, 303–310.
[10] Barbier-Brygoo, H., Vinauger, M., Colcombet, J., Ephritikhine, G., Frachisse, J.-M. and Maurel, C. (1999) Biochem. Biophys. Acta 1465, 199–218.
[11] Nichols, C.G. and Lopatin, A.N. (1997) Annu. Rev. Physiol. 59, 171–191.
[12] Reimann, F. and Ashcroft, F.M. (1999) Curr. Opin. Cell Biol. 11, 503–508.
[13] Ketchum, K.A., Joiner, W.J., Sellers, A.J., Kaczmarek, L.K. and Goldstein, S.A.N. (1995) Nature 376, 690–695.
[14] Lesage, F., Guillemare, E., Fink, M., Duprat, F., Lazdunski, M., Romey, G. and Barhanin, J. (1996) J. Biol. Chem. 271, 4183–4187.
[15] Lesage, F., Guillemare, E., Fink, M., Duprat, F., Lazdunski, M., Romey, G. and Barhanin, J. (1996) EMBO J. 15, 1004–1011.
[16] Derst, C. and Karschin, A. (1998) J. Exp. Biol. 201, 2791–2799.
[17] North, R.A. (1996) Curr. Opin. Cell Biol. 8, 474–483.
[18] Bland-Ward, P.A. and Humphrey, P.P. (2000) J. Auton. Nerv. Sys. 81, 146–151.
[19] Benos, D.J. and Stanton, B.A. (1999) J. Physiol. 520, 631–644.
[20] Maurer, J.A., Elmore, D.E., Lester, H.A. and Dougherty, D.A. (2000) J. Biol. Chem. 275, 22238–22244.
[21] Bassilana, F., Champigny, G., Waldmann, R., de Weille, J.R., Heurteux, C. and Lazdunski, M. (1997) J. Biol. Chem. 272, 28819–28822.
[22] Caterall, W.A. (2000) Neuron 26, 13–25.
[23] Brake, A.J., Wagenbach, M.J. and Julius, D. (1994) Nature 371, 519–523.
[24] Jegla, T., Grigoriev, N., Gallin, W.J., Salkoff, L. and Spencer, A.N. (1995) J. Neurosci. 15, 7989–7999.
[25] Plummer, N.W. and Meisler, M.H. (1999) Genomics 57, 323–331.
[26] Bönigk, W., Müller, F., Middendorff, R., Weynard, I. and Kaup, U.B. (1996) J. Neurosci. 16, 7458–7468.
[27] Wenthold, R.J., Yokotani, N. and Doi, K,Wada,K. (1992) J. Biol. Chem. 267, 501–507.
[28] Doyle, D.A., Morais Cabral, J., Pfuetzner, R.A., Kuo, A., Gulbis, J.M., Cohen, S.L., Chait, B.T. and MacKinnon, R. (1998) Science 280, 69–77.
[29] Doyle, J.L. and Stubbs, L. (1998) Trends Genet. 14, 92–98.
[30] Enright, A.J. and Ouzounis, C.A. (2000) Bioinformatics 16, 451–457.
[31] Karplus, K., Barrett, C. and Hughey, R. (1998) Bioinformatics 14, 846–856.

[32] Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Nucleic Acids Res. 25, 3389–3402.
[33] Promponas, V.J., Enright, A.J., Tsoka, S., Kreil, D.P., Leroy, C., Hamodrakas, S., Sander, C. and Ouzounis, C.A. (2000) Bioinformatics 16, 915–922.
[34] Enright, A.J. and Ouzounis, C.A. (2001) Bioinformatics 17, 853–854.
[35] Ishibashi, K., Suzuki, M. and Imai, M. (2000) Biochem. Biophys. Res. Commun. 270, 370–376.
[36] Bateman, A., Birney, E., Durbin, R., Eddy, S.R., Howe, K.L. and Sonnhammer, E.L. (2000) Nucleic Acids Res. 28, 263–266.
[37] Warmke, J.W. and Ganetzky, B. (1994) Proc. Natl. Acad. Sci. USA 91, 3438–3442.
[38] Schwartz, J.R. and Bauer, C.K. (1999) News Physiol. Sci. 14, 135–142.
[39] Chen, G.-Q., Cui, C., Mayer, M.L. and Gouaux, E. (1999) Nature 402, 817–821.
[40] Wo, Z.G. and Oswald, R.E. (1995) Trends Neurosci. 18, 161–167.
[41] Bargmann, C.I. (1998) Science 282, 2028–2033.
[42] Holm, L. and Sander, C. (1998) Bioinformatics 14, 423–429.
[43] Jan, L.Y. and Jan, Y.N. (1997) Annu. Rev. Neurosci. 20, 91–123.